References

[1] G. Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *J. Alg.*, 55(1): 58–75, 2005.

[2] G. Cormode, H. J. Karloff, and A. Wirth. Set cover algorithms for very large datasets. In *Proc. CIKM*, pages 479–488, 2010.

[3] S. Gog, T. Beller, A. Moffat, and M. Petri. From theory to practice: Plug and play with succinct data structures. In *Proc. SEA*, pages 326–337, 2014.

[4] C. Hoobin, S. J. Puglisi, and J. Zobel. Sample selection for dictionary-based corpus compression. In *Proc. SIGIR*, pages 1137–1138, 2011.

[5] C. Hoobin, S. J. Puglisi, and J. Zobel. Relative Lempel-Ziv factorization for efficient storage and retrieval of web collections. *PVLDB*, 5(3):265–273, 2011.

[6] C. Hoobin, S. J. Puglisi, and J. Zobel. Sample selection for dictionary-based corpus compression. In *Proc. SIGIR*, pages 1137–1138, 2011.

[7] R. M. Karp and M. O. Rabin. Efficient randomized pattern-matching algorithms. *IBM J. Res. Dev.*, 31(2):249–260, 1987.

[8] S. Kuruppu, S. J. Puglisi, and J. Zobel. Relative Lempel-Ziv compression of genomes for large-scale storage and retrieval. In *Proc. SPIRE*, pages 201–206, 2010.

[9] K.-H. Li. Reservoir-sampling algorithms of time complexity $O(n(1+\log(N/n)))$. *ACM Trans. Math. Soft.*, 20(4):481–493, 1994.

[10] C. L. Lim, A. Moffat, and A. Wirth. Lazy and eager approaches for the set cover problem. In *Proc. Aust. Comp. Sc. Conf.*, pages 19–27, 2014.

[11] A. Panconesi and A. Srinivasan. Randomized distributed edge coloring via an extension of the Chernoff-Hoeffding bounds. *SIAM J. Comp.*, 26(2):350–368, 1997.

[12] M. Petri, A. Moffat, P. C. Nagesh, and A. Wirth. Access time tradeoffs in archive compression. In *Proc. Asia Info. Retri. Soc. Conf.*, pages 15–28, 2015.

[13] B. Saha and L. Getoor. On maximum coverage in the streaming model & application to multi-topic blog-watch. In *Proc. SIAM Conf. Data Min.*, pages 697–708, 2009.

[14] J. A. Storer. NP-completeness results concerning data compression. Technical Report 234, Princeton University. Computer Sciences Laboratory, 1977.

[15] J. A. Storer and T. G. Szymanski. Data compression via textual substitution. *J. ACM*, 29(4):928–951, 1982.

[16] J. Tong, A. Wirth, and J. Zobel. Compact auxiliary dictionaries for incremental compression of large repositories. In *Proc. CIKM*, pages 1629–1638, 2014.

[17] J. Tong, A. Wirth, and J. Zobel. Principled dictionary pruning for low-memory corpus compression. In *Proc. SIGIR*, pages 283–292, 2014.

[18] J. S. Vitter. Random sampling with a reservoir. *ACM Trans. Math. Soft.*, 11(1):37–57, 1985.

[19] I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images, Second Edition*. Morgan Kaufmann, 1999.

[20] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Trans. Inf. Th.*, IT-23(3):337–343, 1977.

[21] J. Ziv and A. Lempel. Compression of individual sequences via variable rate coding. *IEEE Trans. Inf. Th.*, IT-24(5):530–536, 1978.